CS 484: Final Project

Due: Dec 13, 2020, by 11:59 PM on Gradescope

- You may work in teams of 2 for this project.
- The final deliverable is twofold:
  1. A (maximum) 4-page PDF file prepared using the ACM template (Word or LaTeX) available at `http://www.acm.org/publications/authors/submissions`. Note that your work will be judged on the quality of the writeup you submit! While we may ask to see your code in case of questions, you are not expected to submit it unless requested. Therefore **the quality of the writeup becomes even more important.**
  2. A (maximum) 5-minute video presentation of your project. Post a link to this video (which you should keep accessible through at least Dec 31) to Piazza (either private or public) by the deadline. You must also include the link at the end of your project report submission.

1 Project Description

The idea of the project is to come up with an interesting *question* and an interesting *dataset*. You may foreground either the question or the dataset, but it must be a question that can be reasonably asked using that data.

For datasets, feel free to browse Kaggle (`https://www.kaggle.com/`) or the UCI ML repository (`https://archive.ics.uci.edu/ml/index.php`). However, you may also collect a dataset of interest to you and make that the primary focus of the project (that is, you could apply the tools we've learned to answer a standard question on the novel dataset). Here are some examples of questions you might be interesed in answering. Remember that these are just ideas, and you should feel free to modify them or come up with your own.

- How do different classifiers compare on a problem with different misclassification costs (*cost-sensitive learning*), and how would this change as the costs for false positive and false negatives changed? For classifiers that output a confidence or probability score, can you optimize the threshold for cost-sensitive classification using cross-validation, and how well does that work?
- Can you collect a novel dataset related to a topic of current interest (e.g. elections, Covid, something else) using the Twitter API and then answer an interesting question (e.g. how different countries / areas / states respond to something?)
- How do different classifiers compare in terms of different fairness metrics on some classic datasets with sensitive or protected features like race or gender?

We encourage you to post your idea to Piazza as a private note, with thoughts on both the question and the data you're interested in, and we will try and give you feedback on the idea ASAP.